



# AI Ethics

Francesca Rossi

IBM AI Ethics Global Leader

# What is AI?

Intelligence and  
rationality

*A scientific discipline that aims to create machines (hw/sw) that show a behavior that would be called intelligent if seen in a human being*

*Rationality*

*Given a problem, to know (and act) how to best solve it*

# Narrow and general AI

## *Narrow AI*

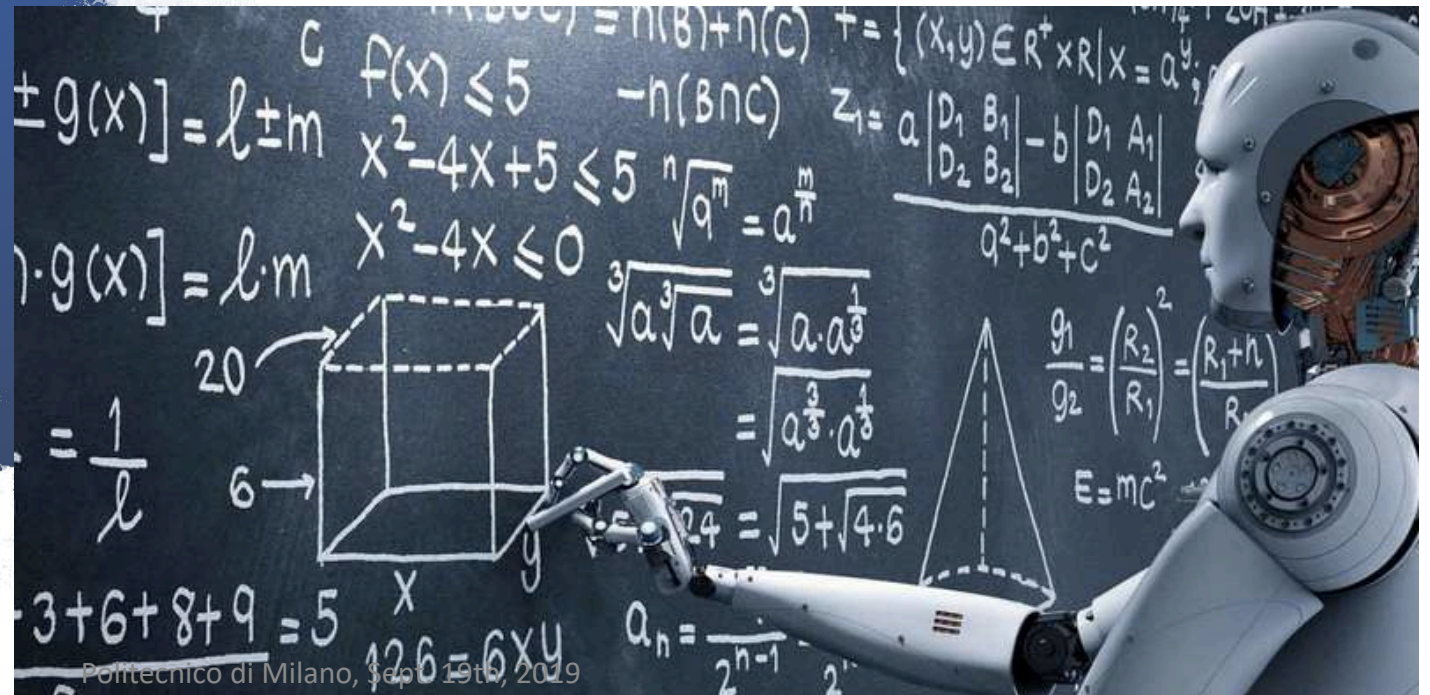
*Can solve specific problems*

*Vertical and specific*

## *General AI*

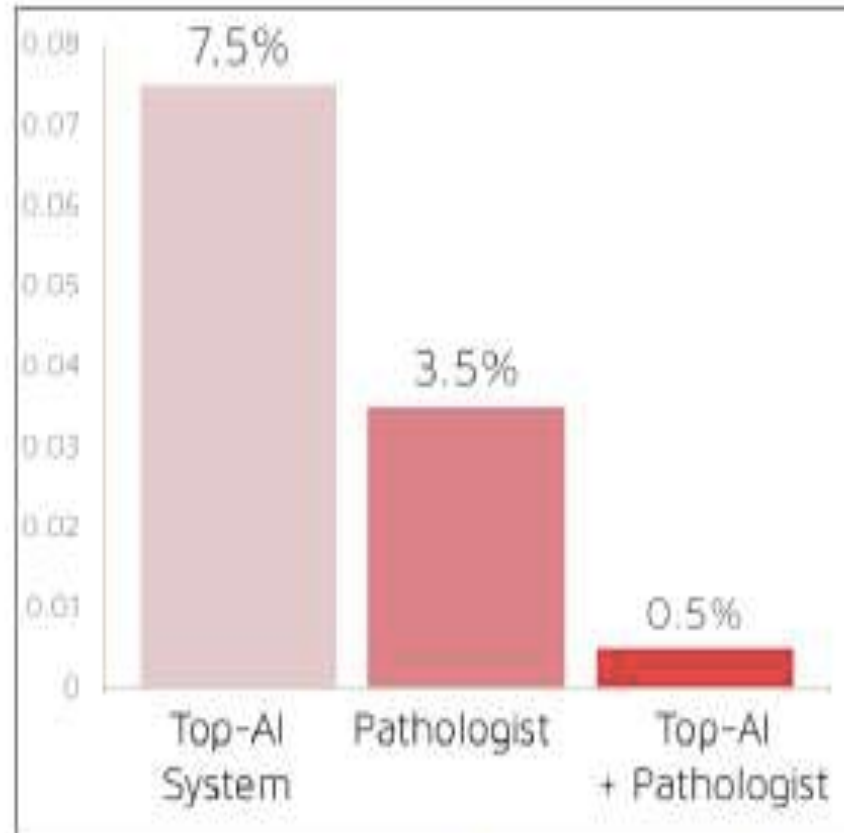
*Can handle different problems and scenarios*

*Horizontal*




# Artificial Intelligence vs Augmented Intelligence

## Human-machine collaboration Complementarity



AI significantly reduces pathologist error rate in the identification of metastatic breast cancer from sentinel lymph node biopsies.



How does it  
work?

How to teach a machine how to solve a  
problem?

1. Logical reasoning

We tell it the steps to be done to solve  
the problem

2. Machine learning

We give examples of problem's  
solutions and we provide methods to  
generalize

# AI, Machine Learning, Deep Learning

## ARTIFICIAL INTELLIGENCE

Intelligent algorithms defined and coded by people into machines



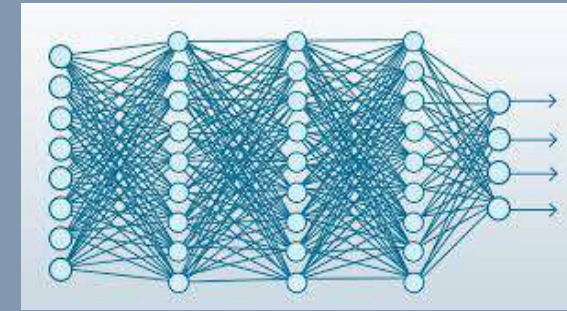
## MACHINE LEARNING

Ability to learn without being explicitly programmed



## DEEP LEARNING

Learning based on Deep Neural Networks



1950's

1960's

1970's

1980's

1990's

2000's

2006's

2010's

2012's

2017's

# Machine/Deep Learning explosion

## YouTube

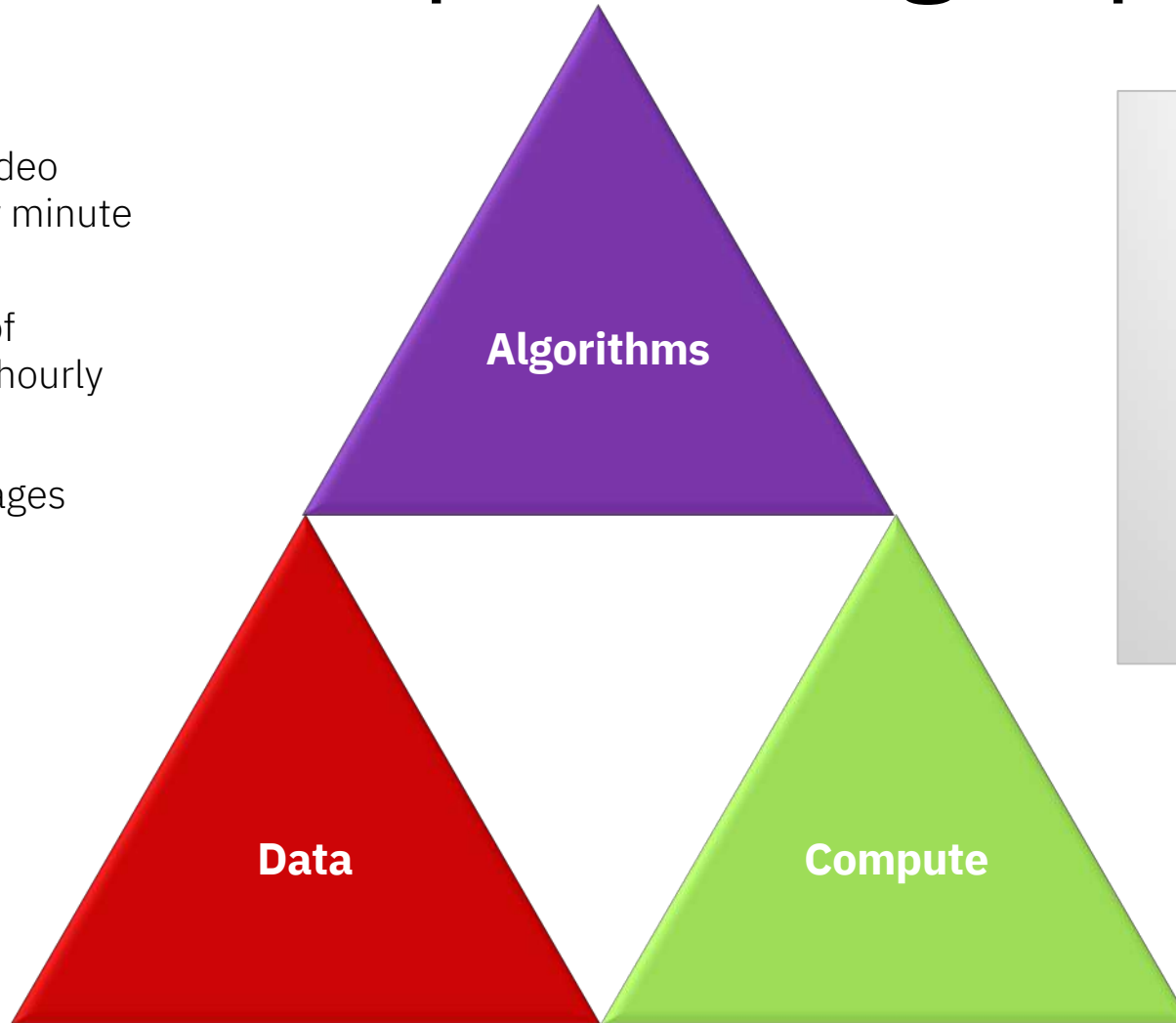
400 hours of video  
uploaded every minute

## Walmart

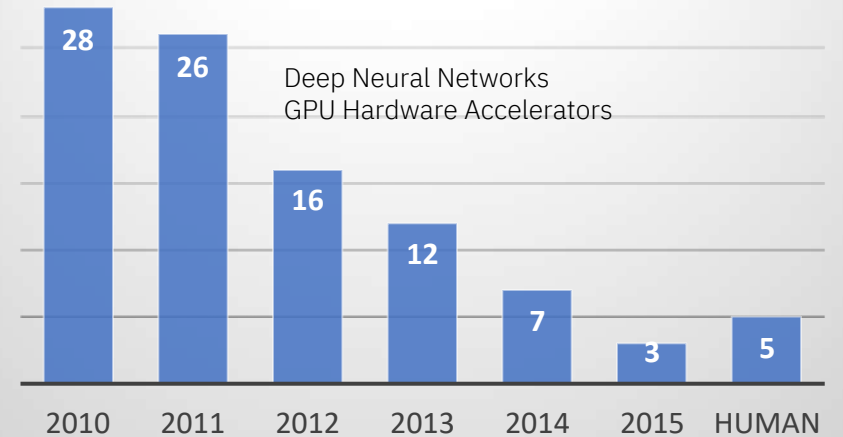
2.5 petabytes of  
customer data hourly

## Facebook

350 million images  
uploaded daily



ImageNet Classification Error



# AI in our life



**amazon.com** Recommended for You

Amazon.com has new recommendations for you based on [items](#) you purchased or told us you own.

 Google Apps Deciphered: Compute in the Cloud to Streamline Your Desktop	 Google Apps Administrator Guide: A Private-Label Web Workspace	 Googlepedia: The Ultimate Google Resource (3rd Edition)
--	---	---



**Google translate**

From:  To:



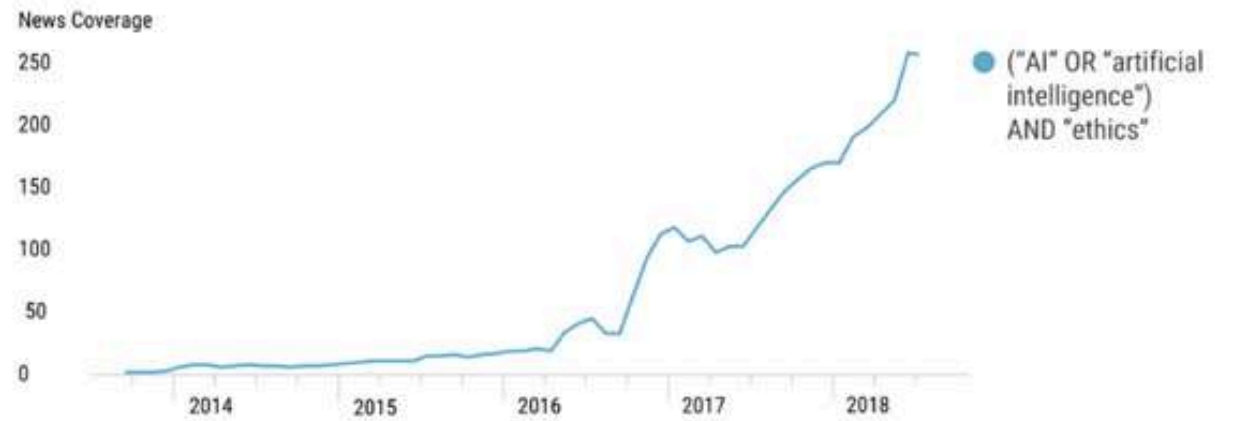


# Current limitations

- Common sense reasoning
- Combination of learning and reasoning
- Natural language understanding
- Learning from few examples
- Learning general concepts
- Robustness/adversarial examples

# AI ethics

Quarterly news mentions of ("AI OR artificial intelligence") AND "ethics" 2014 – Q3 2018



source: cbinsights.com

CBINSIGHTS

# Main concerns

Bias

Careful with the examples!

Value alignment

Ethical, moral, social, and legal constraints

Black box

Must be able to explain its decisions

properties  
of the  
technology

Data issues

Privacy, storage, ownership, use

Accountability

Who is responsible if something goes wrong?

governance  
and rules

Impact on jobs

How to cope with job transformation?

Impact on society

People-machine and people-people interactions

Deep fake

AI can generate content that looks real but it is not

societal  
issues

Autonomous weapons, surveillance systems

Are these acceptable uses?

possible vs desirable  
uses

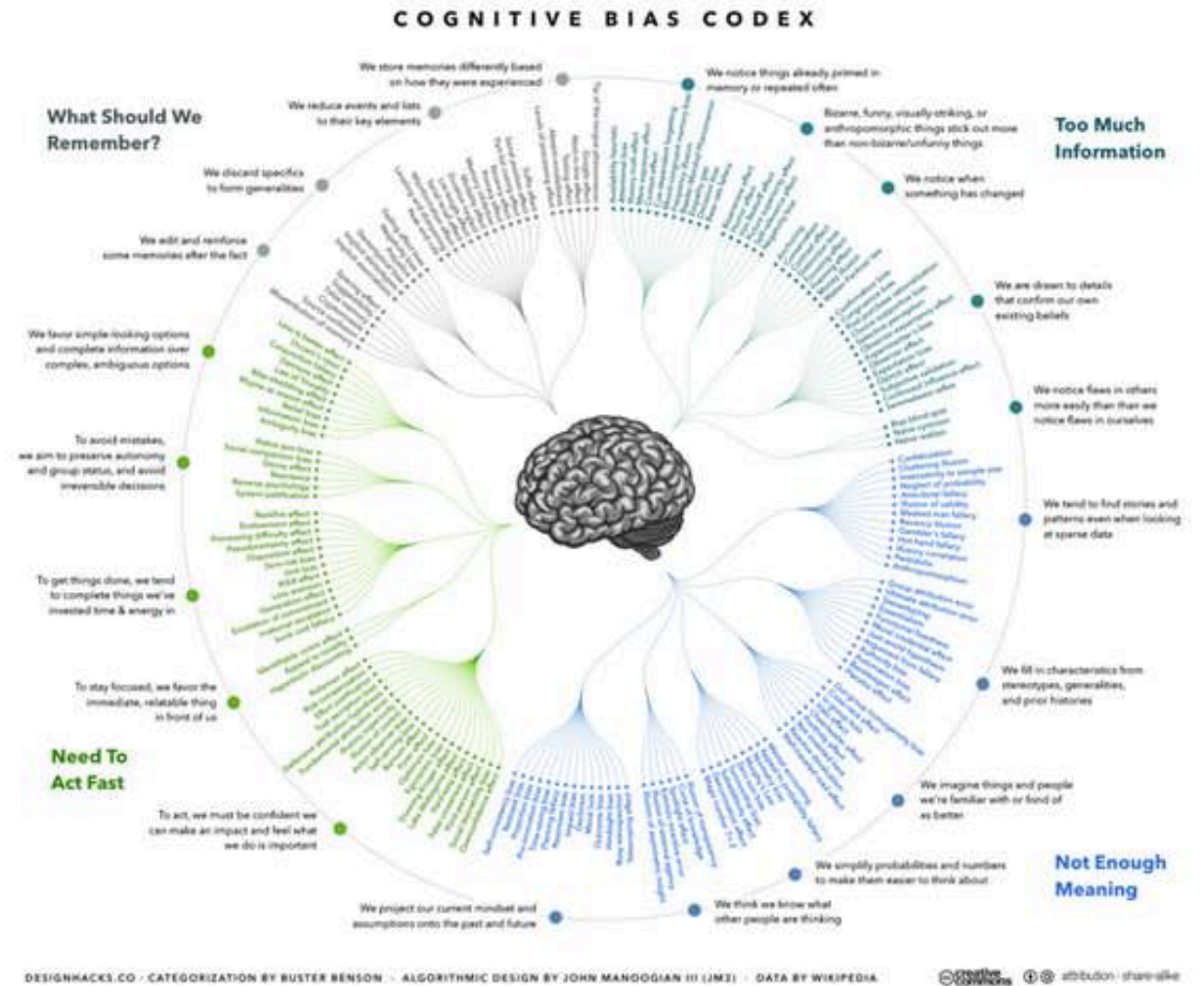
Superintelligence

Will we lose control?

long term concerns

AI bias


Humans are bias



# AI bias


From biased training data  
to unfair decisions

English → Turkish

He is a nurse. She is a doctor.  O bir hemşire. O bir doktor.

English to Turkish

Turkish → English

O bir hemşire. O bir doktor.  She is a nurse. He is a doctor.

Turkish to English

# What is being done?

- **Technological solutions**
  - Bias, explainability, security
- **Other tools**
  - Principles, guidelines, best practices, incentives, standards, certificates, audits, laws, ...
- **Enablers**
  - Education and dissemination
    - AI students, developers, users, impacted communities, policy makers
  - Governance
    - Within each AI company and globally
- **Multi-disciplinary efforts**
  - AI, sociology, psychology, economy, philosophy, law

# What is being done?

## Education

COURSE TITLE	CODE	UNIVERSITY	DEPARTMENT
Algorithms and Society	EECS 395 and COMMST 395	Northwestern University	Computer Science and Communication Studies
<a href="#">Code and Power</a>	LIS 500	University of Wisconsin - Madison	Information School
Designing Field Experiments at Scale	SOC 412	Princeton University	Sociology, Center for IT Policy
Digital Anthropology	ANTH 4020	University of Colorado Boulder	Continuing Education
Ethical and Policy Dimensions of Information, Technology, and New Media	INFO 4601/5601	University of Colorado Boulder	Information Science
Ethical and Social Implications of Data		Marquette University	Computer Science
Ethics in Business Analytics	ITAO 40510	University of Notre Dame	IT, Analytics, & Operations
Ethics in Data Science		University of Utah	Computer Science

Crowdsourced list of AI/CS ethics courses (238 entries so far), Casey Fiesler, CU Boulder

# AI Ethics at IBM

## Adversarial Robustness Toolbox

## AI Explainability 360

### AI Fairness 360

IBM Research Trustworthy AI

Home Demo Resources

AI Fairness 360 Open Source Toolkit

This extensible open source toolkit can help you examine, report, and mitigate discrimination and bias in machine learning models throughout the AI application lifecycle. Containing eight state-of-the-art bias mitigation algorithms developed by the research community, it is designed to translate algorithmic research from the lab into the actual practice of domains as wide-ranging as finance, human capital management, healthcare, and education. We invite you to use it and improve it.

API Docs Get Code

Not sure what to do first? Start here!

- Read More**  
Learn more about fairness and bias mitigation concepts, terminology, and tools before you begin.
- Try a Web Demo**  
Step through the process of checking and remediating bias in an interactive web demo that shows a sample of capabilities available in this toolkit.
- Watch a Video**  
Watch a video to learn more about AI Fairness 360.
- Read a paper**  
Read a paper describing how we designed AI Fairness 360.
- Use Tutorials**  
Step through a set of in-depth examples that introduces developers to code that checks and mitigates bias in different industry and application domains.

Learn how to put this toolkit to work for your application or industry problem. Try these tutorials.

- Credit Scoring**  
See how to detect and mitigate age bias in predictions of creditworthiness using the German Credit dataset.
- Medical Expenditure**  
See how to detect and mitigate racial bias in a care management scenario using Medical Expenditure Panel Survey data.
- Gender Bias in Face Images**  
See how to detect and mitigate bias in automatic gender classification of face images.

### Welcome to the Adversarial Robustness Toolbox

This is a library dedicated to **adversarial machine learning**. Its purpose is to allow rapid crafting, analysis of attacks and defense methods for machine learning models. The Adversarial Robustness Toolbox provides an implementation for many state-of-the-art methods for attacking and defending classifiers. The code can be found on [GitHub](#).

The library is still under development. Feedback, bug reports and extensions are highly appreciated.

### Supported Attacks, Defences and Metrics

The Adversarial Robustness Toolbox contains implementations of the following evasion attacks:

- DeepFool (Moosavi-Dezfooli et al., 2015)
- Fast gradient method (Goodfellow et al., 2014)
- Basic iterative method (Kurakin et al., 2016)
- Projected gradient descent (Madry et al., 2017)
- Jacobian saliency map (Papernot et al., 2016)
- Universal perturbation (Moosavi-Dezfooli et al., 2016)
- Virtual adversarial method (Miyato et al., 2015)
- C&W L<sub>2</sub> and L<sub>inf</sub> attacks (Carlini and Wagner, 2016)
- NewtonFool (Jang et al., 2017)
- Elastic net attack (Chen et al., 2017a)
- Spatial transformations attack (Engstrom et al., 2017)

### AI Explainability 360 Open Source Toolkit

This extensible open source toolkit can help you comprehend how machine learning models predict labels by various means throughout the AI application lifecycle. Containing eight state-of-the-art algorithms for interpretable machine learning as well as metrics for explainability, it is designed to translate algorithmic research from the lab into the actual practice of domains as wide-ranging as finance, human capital management, healthcare, and education. We invite you to use it and improve it.

API Docs Get Code

Not sure what to do first? Start here!

- Read More**  
Learn more about explainability concepts, terminology, and tools before you begin.
- Try a Web Demo**  
Step through the process of explaining models to consumers with different personas in an interactive web demo that shows a sample of capabilities available in this toolkit.
- Use Tutorials**  
Step through a set of in-depth examples that introduce developers to code that explains data and models in different industry and application domains.
- Ask a Question**  
Join our AI Explainability 360 Slack Channel to ask questions, make comments, and tell stories about how you use the toolkit.

**Contribute**  
You can add new algorithms

## 2016 IBM white paper

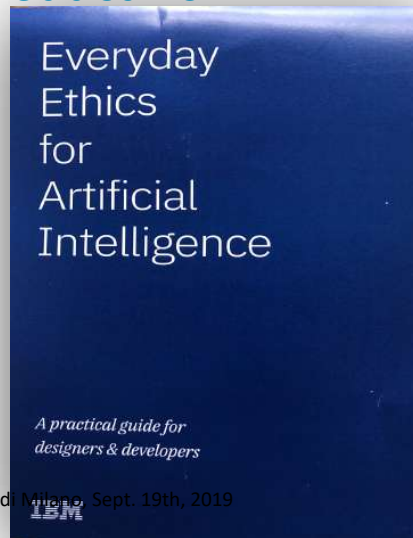


## Science for Social Good

Applying artificial intelligence, cloud and deep science to scale social impact.

## IBM's Principles for Trust and Transparency

## AI ethics developer education



## AI factsheets

- What is the **intended use** of the service output?
- What **algorithms** or techniques does this service implement?
- Which datasets was the service **tested** on?
- Describe the **testing methodology** and **test results**.
- Are you aware of possible examples of **bias**, **ethical** issues, or other **safety risks** as a result of using the service?
- Are the service outputs **explainable** and/or interpretable?
- For each dataset used by the service:
  - Was the dataset checked for **bias**?
  - What efforts were made to ensure that it is **fair** and **representative**?
  - Does the service implement and perform any **bias detection and remediation**?
- What is the **expected performance** on unseen data or data with different distributions?
- Was the service checked for **robustness against adversarial attacks**?
- When were the models last updated?



## Asilomar AI principles

### RESEARCH

1. Research goal
2. Research funding
3. Science-policy link
4. Research culture
5. Race avoidance

### ETHICS AND VALUES

6. Safety
7. Failure transparency
8. Judicial transparency
9. Responsibility
10. Value alignment
11. Human values
12. Personal privacy
13. Liberty and privacy
14. Shared benefit
15. Shared prosperity
16. Human control
17. Non-subversion
18. AI arms race

### LONGER-TERM ISSUES

19. Capability caution
20. Importance
21. Risks
22. Recursive self-improvement
23. Common good



## WORLD ECONOMIC FORUM

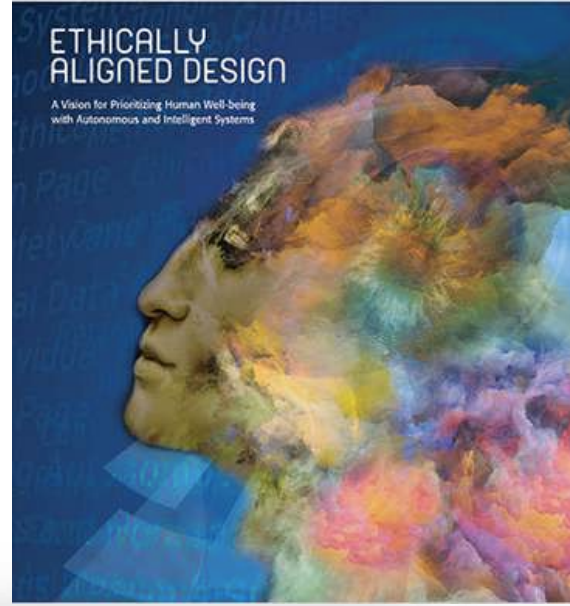


Version II - For Public Discussion



## ETHICALLY ALIGNED DESIGN

A Vision for Prioritizing Human Well-Being with Autonomous and Intelligent Systems



## Partnership on AI

to benefit people and society

One organization

to develop and share the best practices for using and developing AI technologies and providing a global platform to discuss how AI will influence people and society.

7 Thematic Pillars



- Safety Critical AI
- Fair, Transparent, and Accountable AI
- AI, Labour and the Economy
- Collaborations between People and AI systems
- AI and Social Good
- Social and Societal Influences of AI
- Special Initiatives



30+ Partners



## AI for Good Global Summit

An ITU experience



LEVERHULME CENTRE FOR THE FUTURE OF INTELLIGENCE



## AAAI / ACM conference on ARTIFICIAL INTELLIGENCE, ETHICS, AND SOCIETY

## Everyday Ethics for Artificial Intelligence

A practical guide for designers & developers



INDEPENDENT HIGH-LEVEL EXPERT GROUP ON ARTIFICIAL INTELLIGENCE SET UP BY THE EUROPEAN COMMISSION



## ETHICS GUIDELINES FOR TRUSTWORTHY AI

INDEPENDENT HIGH-LEVEL EXPERT GROUP ON ARTIFICIAL INTELLIGENCE SET UP BY THE EUROPEAN COMMISSION



## POLICY AND INVESTMENT RECOMMENDATIONS FOR TRUSTWORTHY AI

# Embedding ethics into decision support systems

Personalization vs (ethical) behavioral constraints

Preferences are essential to allow for personalized services  
Online recommendations, healthcare, financial advisors, etc

Also a way to tell a system what we want from it

But they need to be combined with other priorities to avoid undesired actions

Ethics principles, moral values, business guidelines, behavioral constraints, common sense reasoning





# Value alignment

To achieve a goal, given by a human  
in the best way

possibly being creative and innovative  
while being aligned to the appropriate values for  
the task

# Ethically bounded AI

- How do we bound the behavior of autonomous agents, without explicitly telling them what to do, in a way that it will achieve the goal while complying with appropriate ethical/behavioral constraints?
- Two main approaches:
  - Top Down: write down all the rules and have the agent follow them
    - We need to know the best strategy to solve the problem
  - Bottom Up: show the agent appropriate actions
    - Data-driven approach

# Two explored solutions

## 1. Recommendation systems

- Goal: to teach AI systems how to obey behavioral constraints learned by observation while still being responsive to the feedback from users
  - Reinforcement Learning approach
  - Examples to describe the ethical constraints, learnt offline
  - Constrained RL behavior during online use

## 2. Preferences and ethical priorities

- Goal: To achieve personalization while not compromising essential values and principles
  - Preference frameworks (CP-nets) to model both preferences and ethical guidelines
  - Distance between CP-net structures
  - Distance thresholds to decide if agent can follow its preferences or must be better aligned to ethical priorities











Conference papers: AAI 2018, AAMAS 2018, AIES 2018

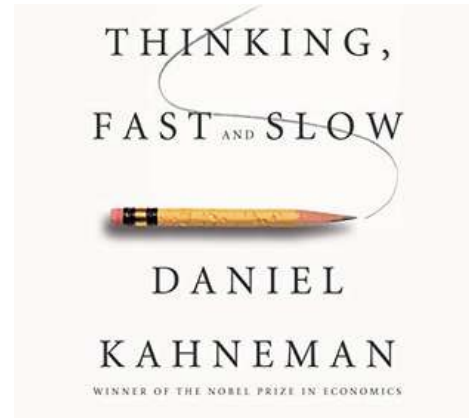
Book chapters:

- *Artificial Intelligence Safety and Security, CRC Press, 2018*
- *Ethics of AI, Oxford University Press, 2019*

# Which is better?

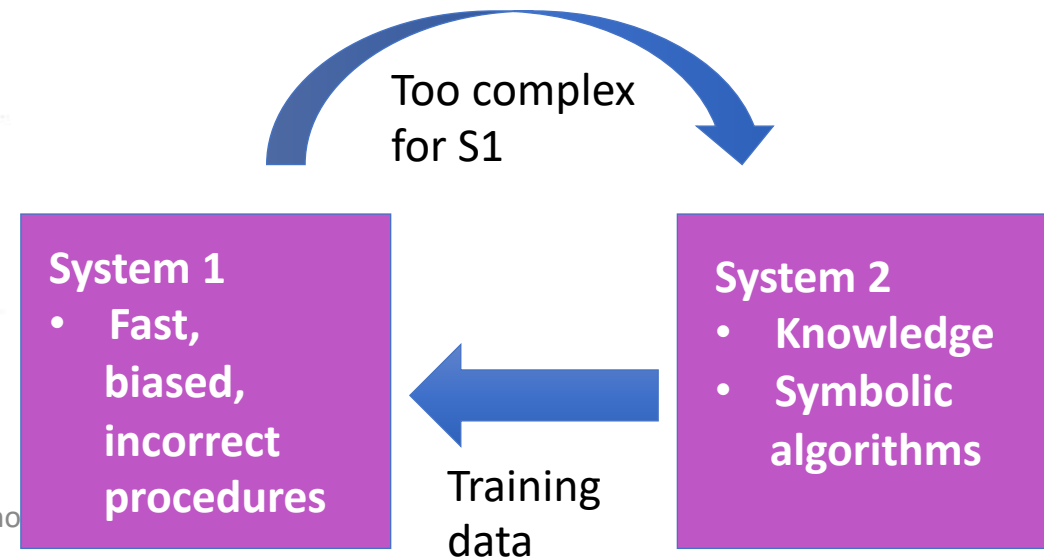
Should we choose?  
What happens in our brain?

System 1	System 2
 Fast	 Slow
 Unconscious	 Conscious
 Automatic	 Effortful
 Everyday Decisions	 Complex Decisions
 Error prone	 Reliable



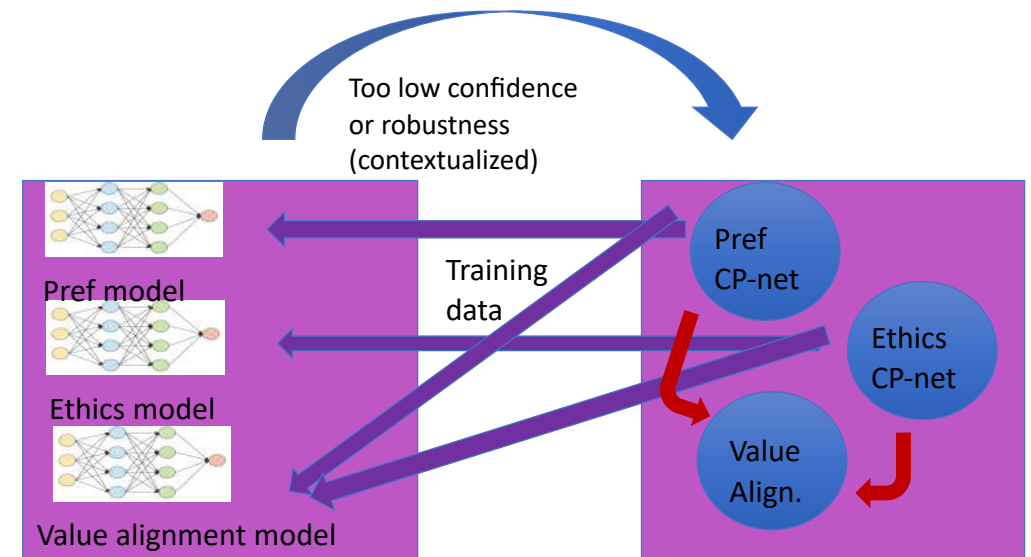
# Fast and slow AI

- System 1 resembles a data-driven approach
  - Training data provided over time by System 2
  - After a while, some tasks pass to system 1, while others always require system 2
  - Ex.: reading a word, arithmetic 2-digit multiplication
- System 2 resembles a symbolic/logic approach
  - Understands how to tackle new or computationally difficult tasks
  - Computational complexity is (one of) the trigger(s) for system 2



# Fast and slow preferences and ethical constraints

- A system 1 and a system 2 version for both preferences and ethical constraints
- The Pref/Ethics/VA modes answer to
  - What is the most preferred/ethical choice?
  - Is choice A dominated by choice B?
  - How far are my preferences from the ethical constraints?
- But only after he got enough training data from the Pref/Ethics/VA symbolic/logic procedure





# What next?

- Implement the full dual-agent architecture
- Understand when S1 should call S2
  - or when S2 should awake and override S1
- Contextualize to tasks and scenarios
- Where to place causality and common sense reasoning?
- More than one ethical theory
  - Deontology: constraints and priorities
  - Utilitarianism
  - Contractualism
- Elephant in the Room: Where do values and ethical constraints come from?
  - Multi-stakeholder approach





Thanks!